The State of Data Curation at NeurIPS: An Assessment of Dataset Development Practices in the Datasets and Benchmarks Track

Eshta Bhardwaj[1,3], Harshit Gujral[2], Siyi Wu[2], Ciara Zogheib[1], Tegan Maharaj[1], and Christoph Becker[1,3]

[1]Faculty of Information, University of Toronto
[2]Department of Computer Science, University of Toronto
[3]Digital Curation Institute, University of Toronto

Abstract

Data curation is a field with origins in librarianship and archives, whose scholarship and thinking on data issues go back centuries, if not millennia. The field of machine learning is increasingly observing the importance of data curation to the advancement of both applications and fundamental understanding of machine learning models – evidenced not least by the creation of the Datasets and Benchmarks track itself. This work provides an analysis of recent dataset development practices at NeurIPS through the lens of data curation. We present an evaluation framework for dataset documentation, consisting of a rubric and toolkit developed through a thorough literature review of data curation principles. We use the framework to systematically assess the strengths and weaknesses in current dataset development practices of 60 datasets published in the NeurIPS Datasets and Benchmarks track from 2021-2023. We summarize key findings and trends. Results indicate greater need for documentation about environmental footprint, ethical considerations, and data management. We suggest targeted strategies and resources to improve documentation in these areas and provide recommendations for the NeurIPS peer-review process that prioritize rigorous data curation in ML. We also provide guidelines for dataset developers on the use of our rubric as a standalone tool. Finally, we provide results in the format of a dataset that showcases aspects of recommended data curation practices. Our rubric and results are of interest for improving data curation practices broadly in the field of ML as well as to data curation and science and technology studies scholars studying practices in ML. Our aim is to support continued improvement in interdisciplinary research on dataset practices, ultimately improving the reusability and reproducibility of new datasets and benchmarks, enabling standardized and informed human oversight, and strengthening the foundation of rigorous and responsible ML research.

1 Introduction

The NeurIPS Datasets and Benchmarks (D&B) track was created in 2021 to enhance the development of datasets in line with the exponential growth of applications of machine learning (ML). A key challenge this track aimed to address is that of datasets being used outside their original scope as benchmarks [43, 81, 87, 102] – among other potential issues, this creates the possibility of field-level overfitting, as well as unanticipated ethical and privacy problems. NeurIPS sought to address this by encouraging the development of new datasets for ML, improving the quality of datasets being produced, and emphasizing the importance of the role of data within ML. The introduction of new, tailored peer-review guidelines also enabled and incentivized publication of datasets and benchmarks. The D&B track is thus uniquely positioned to influence and guide the quality and ethicality of datasets being released and bolster responsible dataset development practices.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

Recent research in fairness, accountability, and transparency has proposed to reduce bias in models, via datasets, through the improvement of dataset documentation [29, 50, 52, 56, 61, 64, 86, 90]. Particularly, recent research on ethical data curation for ML datasets [47] emphasizes the adoption of concepts and processes from library and archival studies and digital curation in order to improve the documentation of dataset contents and data design decisions [47, 19, 39, 97]. **Data curation**, a subset of digital curation and information science, is "...the activity of managing data throughout its life cycle; appropriately maintaining its integrity and authenticity; ensuring that it is properly appraised, selected, securely stored, and made accessible; and supporting its usability in subsequent technology environments." [75, p. 203]. We recently translated data curation concepts for the ML dataset development context [11]. The results enable us to apply data curation to the field of ML so that dataset development practices can be evaluated and improved.

Contributions. Our goal is to document and improve the standard of dataset development in NeurIPS so that future benchmarks and datasets can be effectively found, easily accessed, ethically used, consistently evaluated, and appropriately reused. To these ends, we present a systematic dataset documentation evaluation framework to organize the assessment of curation practices for ML dataset development. The framework is composed of a rubric and toolkit developed through an iterative multi-stage process to arrive at the rubric elements relevant for evaluation and corresponding criteria for their assessment, as well as supplementary material packaged as a toolkit to aid in the application of the rubric. Here, we establish the feasibility of this framework as an auditing tool for dataset documentation by performing a systematic evaluation of a sample set of 60 datasets published in the NeurIPS D&B track between 2021-2023. We analyze the assessments to evaluate how data curation is currently performed in ML and show how it can be improved. Our results demonstrate that documentation quality varies widely across datasets and reveal a lack of documentation and reflexivity on environmental footprint, the situatedness and non-neutral nature of data, ethical considerations, and data management. We recommend how to improve this in Section 5 and make a proposal for NeurIPS peer-review changes.

2 Background

2.1 The Importance of Data in Machine Learning

Increasingly, machine learning research has turned towards the improvement of data to improve model results and fundamental understanding. Research areas include the role of data (distribution) in learning theory and generalization [6, 27, 26, 38, 101, 51], explicitly data-centric machine learning [25], the construction of datasets [3, 30, 46, 82, 85, 93], addressing a greater number of areas and problem domains [5, 23, 44, 48, 98, 103], the development of ethical models/frameworks around AI and data [2] within sound and music computing, computer vision, natural language interfaces, and more [7, 17, 24, 58, 92].

NeurIPS has responded to the rising urgency and recognized impact of data through the introduction of the Datasets and Benchmarks (D&B) track [102]. Submissions to the NeurIPS D&B track highlight aspects of data work critical to the development of machine learning datasets. Specifically, [84] looks at impacts of unmanaged citations of dataset derivatives, continued usage of datasets after their retraction, ambiguous dataset documentation, non-restrictive and ineffective licensing, and lack of long-term data stewardship. A data quality framework developed for datathons [62], is also applicable for the examination of data quality of ML datasets as it covers 5 broad quality dimensions. A checklist for ethical data collection of human-centric computer vision datasets highlights that further emphasis should be placed on moving away from general-purpose datasets to clearly-defined dataset collection, providing recommendations for obtaining consent and maintaining privacy, and proposing how to examine, acknowledge, and enhance dataset diversity [4].

In 2023 NeurIPS released a Code of Ethics [65] to supplement the NeurIPS Code of Conduct [74]. The code of ethics includes information on how to report and prevent harms from research that involves human participants, data concerns including privacy, consent, etc., and societal harms such as concerns of safety, security, discrimination, harassment, environment, human rights, and bias. Alongside this, there were ethics guidelines released for reviewers [71] which point reviewers to checklists and a framework for evaluating general ethical conduct, as well specifications for human-related data, data concerns around compliance, consent, and regulations, and negative societal impacts. Guides and further information about how to report on these areas have been released by NeurIPS and others [13, 68]. Since 2021, the D&B track has seen immense growth and success, from 174 of 484 submitted papers accepted in 2021 to 322 of 987 submitted papers accepted in 2023 [67, 73].

As the landscape of data-focussed ML research continues to evolve, there remain open challenges for the field to tackle. For example, a review of AI impact statements in NeurIPS papers from 2021 [53] finds that 'agency' and 'responsibility' are two key themes in the statements whereby dataset creators feel they are not in control of the negative impacts of their work if it was to be misused by malicious users. Authors typically reassign the responsibility to identify and safeguard against potential negative impacts to other practitioners, or state that the potential negative impacts of their work are the same as those that exist for that domain, i.e., further work is needed to recognize contributions as non-neutral and take accountability of how design decisions impact outcomes [53].

2.2 Data Curation

Data curation's influences as a field include information sciences, digital libraries, and archival sciences [78, 88, 1, 49, 16, 39]. It thus has deep-rooted and established methods and discourse on how to maintain large amounts of data and manage ethical concerns. Data curation as a component of digital curation takes a **lifecycle approach** to the management of digital data [75]. Lifecycle approaches divide the digital curation practice into stages, which differ in their details but have in common a broad view including the pre-use, use, and post-use of a dataset. For example, the Digital Curation Center's model specifies 'conceptualize', 'create or receive', 'appraise and select', 'ingest', 'preservation action', 'store', 'access, use, and reuse', and 'transform' [34]. Each stage of curation consists of activities, designated roles and responsibilities, specified technical, legal, ethical, and operational considerations as well as policies that institutionalize the discussed activities and responsible parties [34].

ML studies on data practices ("...what and how data are collected, managed, used, interpreted, reused, deposited, curated, and so on..." [12, p. 55]) also called dataset development and data work most closely resemble examinations of the processes of data curation (despite often using the term to connote data collection, e.g., [56, 50, 35]). Other works in this space discuss the importance of documentation and propose new frameworks for it [8, 18, 29], review intrinsic and extrinsic biases in the dataset development process [64, 79, 63], and highlight the power dynamics involved in data quality, data work, and documentation [60].

In recent work [11], we performed a thorough literature review of data curation practices, translated them to be applied in a machine learning setting, and presented a preliminary analysis of a few NeurIPS D&B track datasets, focussed on the interdisciplinary process of adopting data curation for ML. Particularly, we highlighted the need for tools to translate the standards for transparency and accountability. In contrast to previously mentioned studies, our framework enables the application of *data curation* principles and concepts in practice.

Other authors also point to the need for dataset creators to *actively* document and steward their datasets [84], in contrast to much of the static documentation which is common in ML. These recommendations can be translated into practice-based processes: by seeing dataset development in ML as data curation and adopting its norms and practices, the NeurIPS community and ML research practices broadly can gain an elevated standard of documentation and resulting benefits to model performance, responsibility, and fundamental understanding.

3 Methods

Research Questions. What are the strengths and weaknesses of NeurIPS dataset documentation practices considered through a data curation lens? In other words, how well curated are NeurIPS datasets and benchmarks? To study this, we examine (1) What constitutes a well curated dataset? (2) How feasible is the adoption of data curation principles to assess ML datasets? and (3) What is the state of data curation at NeurIPS and how can it be further advanced?

Approach. We developed an evaluation framework to assess data documentation practices, i.e., curation processes, and applied it to recently published ML datasets in the NeurIPS datasets and benchmarks track. This track was precisely chosen because of its relevance in publishing such contributions but also in influencing the quality of datasets that are accepted. The evaluation framework consists of a rubric used to evaluate dataset documentation and design decisions and a toolkit which supplements the rubric by providing additional information on how to apply the rubric effectively. This framework was applied to manually assess 60 ML datasets in three steps.

1. We established the **initial construction and design** of our evaluation framework consisting of the rubric and toolkit and reviewed the **preliminary feasibility** of the framework by applying it to 25 datasets across 4 rounds [11]. In these rounds, we continued to develop the framework iteratively based on the evaluation results [11]. We reflected and reported on the initial process of designing the framework such as the benefits resulting from the diverse perspectives of an interdisciplinary team, the lessons learned while applying the framework, and how we used the data from the initial application of the rubric to iteratively refine it and yield more consistent evaluations [11].
2. We **examined the consistency in application** by measuring inter-rater reliability (IRR). To claim that our framework is consistent, reliable, and accordingly feasible, we conducted another round of evaluations consisting of 5 datasets (round 4 and disagreement review) with the framework fully developed. We therefore address RQ 1 with our most updated version of the framework and RQ 2 with the final fourth round of evaluations that firmly establishes the framework's reliability through iteratively improving IRR results.
3. We applied this framework to assess 30 additional datasets to **evaluate current practices** of data curation in ML dataset development and areas where improvement was needed (RQ 3).

Evaluation Framework. We grounded our framework in data curation principles, emphasizing documentation, transparency, and ethical considerations. We started with key aspects of data curation relevant to ML and followed with iterative refinement through internal reviews and adjustments to evaluation criteria, guided by digital curation lifecycle models, FAIR data principles, and environmental sustainability and justice considerations. The rubric, provided in the Appendix, consists of 18 elements across five categories. In [11], we presented results from a training round and rounds 1-3. After round 3, we updated and refined the criteria for 13 elements and added additional guidance in the toolkit for interpreting authenticity, reliability, and representativeness. We present the up-to-date version of the framework along with the changes made between versions and their rationale in the Appendix.

The **scope** category has 2 elements, 'context, purpose, motivation' and 'requirements', which emphasize the requirement for a dataset creation plan and addressing intrinsic biases. The **ethicality and reflexivity** category has 4 elements, 'ethicality', 'domain knowledge and data practices', 'context awareness', and 'environmental footprint', covering a range of documentation requirements to increase reflection and accountability in the dataset creation process. The **data pipeline** category includes 'data collection', 'data processing', and 'data annotation', prompting reflection on how and why choices were made and their implications. The **data quality** category contains 'suitability', 'representativeness', 'authenticity', 'reliability', and 'structured documentation', to ensure the consideration of a broad set of qualities that impact how well a dataset can be appropriately and responsibly reused. The **data management** category covers FAIR principles [99] - findability, accessibility, interoperability, and reusability - included to evaluate the transparency of data management considerations. Each rubric element is assessed on minimum standard criteria (with a score of 'pass' or 'fail') that detail the expected level of documentation. Elements that pass the minimum standard are also assessed on a standard of excellence (with a score of 'full', 'partial', or 'none'). Therefore, the conceptualization of the rubric defines what a well-curated dataset must document. The **toolkit** is a supplementary resource that introduces concepts from data curation and serves as a manual to the rubric. It contains instructions and guidance on how to evaluate datasets, how to interpret specific elements, guiding principles, recommendations, FAQ, sample evaluations, a glossary, and further readings. The toolkit is provided in the Appendix.

Iterations. In order for data curation to provide robust norms for ML dataset development, our framework has to *enable consistent use*. To evaluate consistency, we measured inter-rater reliability (IRR) as we iteratively refined the rubric over multiple rounds of evaluation. The preliminary stage of refining the rubric occurred across the first 4 rounds of evaluations (namely training, round 1, round 2, round 3) [11]. Each round involved improvements based on feedback and observations, ensuring the rubric and toolkit were effectively refined and validated. This was structured around several key activities. We began with a training round to help reviewers become acquainted with the rubric and foundational data curation concepts, significantly reducing initial discrepancies and increasing IRR in the upcoming rounds. Following each evaluation round, we gathered feedback from reviewers and identified specific areas of the rubric that needed adjustments to better convey the expectations and reduce ambiguity. We refined definitions, provided clearer examples, and better aligned the rubric elements with practical evaluation scenarios. This established preliminary feasibility.

Consistency. To establish the framework's feasibility and consistency, we performed additional rounds of evaluations. Across training to round 4, three reviewers assessed each of the 30 datasets in a fully crossed design [57] thus we calculated IRR using a two-way mixed, consistent, average-measures intra-class coefficient (ICC) to assess the consistency of the raters' evaluations of rubric elements measured on an ordinal scale across subjects [31]. In rounds 3 and 4, we additionally performed a "disagreement review". Once reviewers had completed their evaluations, they reviewed other evaluations and engaged in a brief discussion period in which they could debate, review, and update their scores and comments. Given the interpretative nature of the framework, this collaborative disagreement review enabled improved understanding of the rubric concepts and mitigated potential errors such as overlooking provided documentation. It also led to greater consistency while simultaneously leveraging the diverse perspectives of reviewers to enhance the richness and accuracy of the dataset evaluations. With the framework's feasibility established, we evaluated additional datasets.

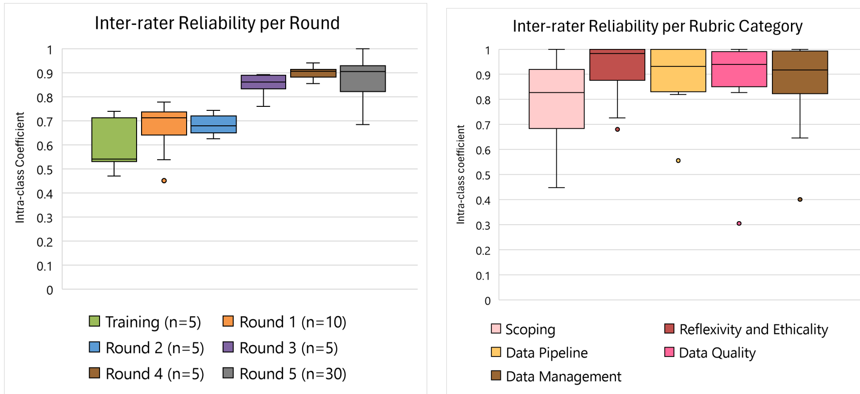
Application. To understand precisely how data curation can contribute to the advancement of ML dataset documentation practices, we performed a final round of evaluations (round 5). In this round, we evaluated 30 datasets published in the NeurIPS D&B track from 2021-2023. A full list of evaluated datasets from all rounds can be found in the Appendix. The datasets were randomly selected after filtering all published papers at the NeurIPS D&B track for dataset contributions. The filtering process is described in the Appendix. In the final round, four reviewers performed double coding for 30 datasets, each reviewing on average 15 datasets, including a disagreement review. Accordingly, we measured IRR with a one-way mixed, consistent, average-measures intra-class coefficient (ICC). After the disagreement review, additional corrections were made for consistency, see Appendix. All 60 dataset evaluations and analysis files can be found hosted on Zenodo [10].

Analysis. We analyze to what extent criteria were fulfilled for 1) each dataset and 2) each rubric element. This enables a review of whether data curation can provide guidance for documentation for NeurIPS datasets and precisely in what capacity that guidance is needed.

4 Results

R1. Inter-rater reliability suggests the evaluations are consistent and reliable. We observed a quantifiable improvement in IRR per dataset across successive evaluation rounds. The ICC values progressively increased moving from "fair" to "excellent" agreement among raters. In the final round, the median ICC value for the 30 datasets evaluated was 0.90 (Figure 1a). Despite the high level of qualitative human interpretation present when evaluating IRR across elements as compared to datasets, the final round had very high agreement, with ICC values with medians ranging from 0.83-0.98 across rubric categories (Figure 1b). The improvements in IRR confirm the effectiveness of our iterative refinement approach. By continuously enhancing the rubric and its guidelines, we achieved a high level of consistency in evaluations, demonstrating the rubric's potential as a reliable tool for assessing dataset documentation in machine learning. This high level of agreement underscores the clarity and effectiveness of the rubric's criteria in guiding evaluators to consistent outcomes. These findings are critical as they establish the rubric's credibility and pave the way for its broader application and acceptance within the ML community. Additionally, the findings demonstrate the utility and rigor of qualitative human evaluations. ICC values for each of 5 rounds is shown in Figure 1a and across rubric categories in Figure 1b. Additional results are provided in the Appendix.

R2. Documentation quality varies widely across datasets. To gauge the extent of documentation provided, we calculated for each dataset the percentage of rubric elements that received a 'pass' and 'fail' score for the minimum standard and a 'full', 'partial', and 'none' score for the standard of excellence. Since each dataset was evaluated by 2 reviewers and the rubric consisted of 18 elements, we averaged the score across both reviewers and divided by 18. The results are shown in Fig 2a and 2b. Across all datasets evaluated during the final round, one fulfilled 86% of the minimum standard criteria (highest pass rate), while another fulfilled only 39% (lowest pass rate), a 47% difference. The absolute difference between the best and worst performing papers at the standard of excellence criteria is similar, with the best-performing paper scoring 'full' on 50% of the standard of excellence and the two worst-performing receiving no 'full' scores. These results demonstrate that documentation varies



(a) IRR for evaluation rounds (b) IRR for rubric categories in round 5

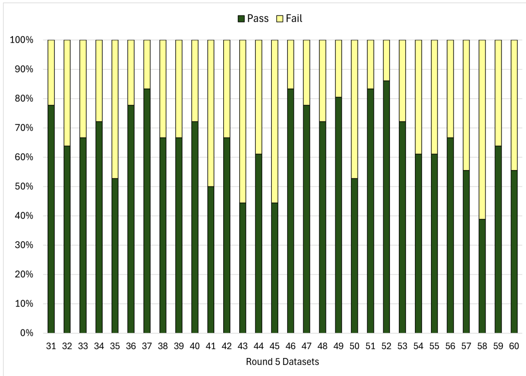
Figure 1: Inter-rater reliability (IRR) **(a)** Across evaluation rounds, and **(b)** Within round 5 across rubric categories. Improvement of IRR across rounds and ultimate high IRR across categories provides evidence that the multi-stage quality and consistency process described in Sec. 3 was successful. In addition to this quantitative measure, we conducted qualitative participatory evaluations with reviewers in each round; see **R1** and Appendix.

widely across datasets and there is great scope for improvement in documentation practices from a data curation lens, particularly to meet a standard of excellence.

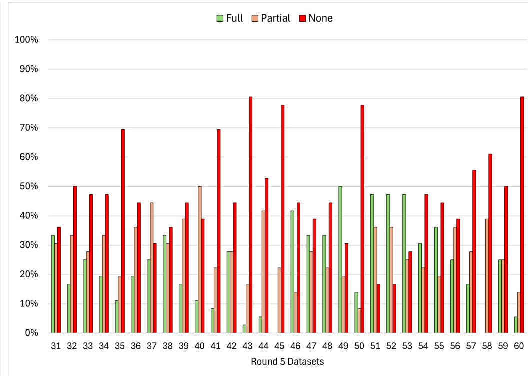
R3. NeurIPS prioritizes model-work adjacent documentation. To analyze where data practices could be improved, we measured the completion of documentation for each rubric element and category by calculating the number of 'pass', 'fail' scores and 'full', 'partial', 'none' scores for all 60 evaluations (2 per each dataset in round 5) and divided the number by 60. Notably, NeurIPS papers tended to perform better at certain rubric elements than at others (see Figure 2c, 2d). Documentation for the minimum standard for 'context, purpose, motivation', 'suitability', and 'reliability' was 100% fulfilled. Additionally, all 3 elements in the data pipeline category were 93% fulfilled. This highlights the areas that are prioritized and considered primary for documentation by dataset creators. These are also areas that are standard to report for publication. NeurIPS has also been able to guide and encourage greater focus through the suggested submission requirements for 'structured documentation' (i.e., datasheets [29], data statements [8], etc.) and some aspects of the data management rubric elements. For example, documentation for a dataset clearly stated the problem domain of NLP and computer vision and the relevance of the new dataset being introduced in creating speech-based rather than text-based input for assistive devices ('context, purpose, motivation'), discussed the feasibility of their dataset ('suitability'), and provided a datasheet ('structured documentation').

R4. Documentation is rarely context-aware and typically does not quantify environmental footprint. The rubric elements with the worst performance across round 5 evaluations are 'context awareness' and 'environmental footprint', both with 0% pass rates of the minimum standard (and subsequently of the standard of excellence). Papers fail the 'context awareness' rubric criteria by not including a dedicated positionality statement (a statement of authors' institutional affiliations is not considered as a statement of or reflection on positionality). For the standard of excellence, less than 20% of papers receive a 'full' or 'partial' score for the 'ethicality' standard of excellence. That is: even those papers that make use of the proportionality principle and document informed consent tend to do so only as much as required by ethics checklists, with additional ethical discussion rarely included. The evaluated datasets also fail the 'environmental footprint' criteria because none of them quantitatively assess the environmental footprint associated with dataset creation.

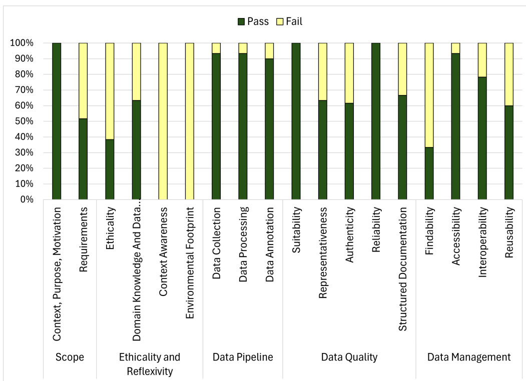
R5. Documentation often remains incomplete. The results indicate that even for those datasets with well-documented elements for the minimum standard, rigorous documentation is ultimately lacking. For example, in the case of 'reliability', papers tend to pass the minimum standard by describing the phenomena represented by the data (e.g., describing the videos from which screen capture data were generated), but fail the standard of excellence by not providing a mechanism by which others could verify what was being represented (e.g., no way for anyone else to access the videos used to produce screen captures). As in the case of 'reliability', and as intended in the rubric's design, papers perform



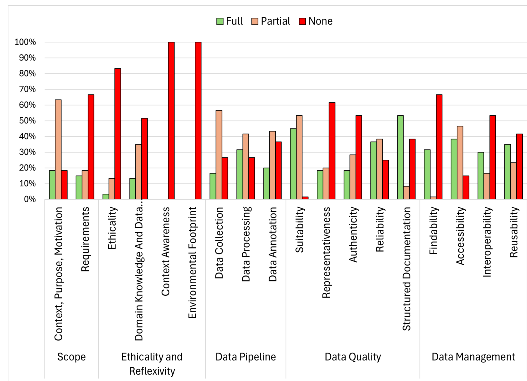
(a) Per dataset, Minimum standard



(b) Per dataset, Standard of excellence



(c) Per element, minimum standard

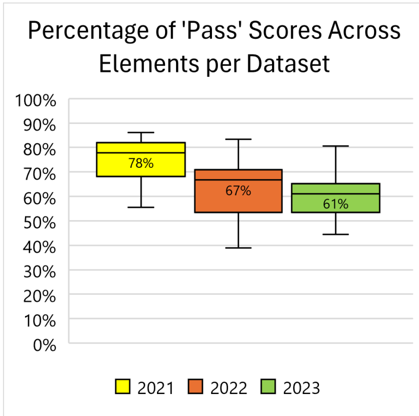


(d) Per element, Standard of excellence

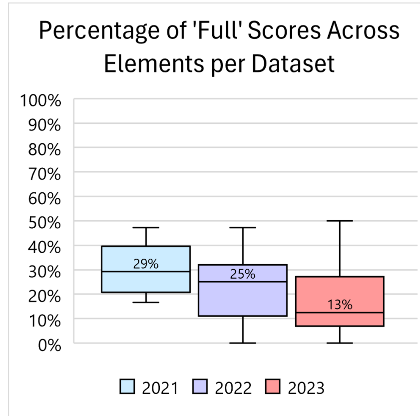
Figure 2: Percentage of completed documentation per dataset (a,b) and per element (c,d) in round 5 (i.e. after a multi-step iterative process to improve quality). In **(a)** we observe that the highest scoring dataset fulfilled 86% of criteria to meet the minimum standard of quality while the lowest fulfilled only 39%; in **(b)** for the standard of excellence we see similar spread (approximately 50% difference but lower attainment (highest fulfilled 50% of criteria, lowest two fulfill none of the criteria for excellence); see **R2**. In both **(c)** minimum standard and **(d)** excellence we observe that those elements more closely related to model-work (such as 'suitability' and 'reliability') are more consistently fulfilled; see **R3**.

better at the minimum standard than at the standard of excellence: 14 out of 18 rubric criteria have a minimum standard pass rate over 50%, compared to 1 of 18 with 'full' scores over 50% and 3 of 18 with 'partial' scores over 50% for the standard of excellence.

R6. Findings suggest no improvements occurred over time. We evaluated an even distribution of datasets published in 2021, 2022, and 2023 for round 5. Figure 3 shows the results of the percentage of 'pass' and 'full' scores across elements for each dataset summarized by year. Particularly, we can observe a slight downward trend in documentation scores across the years evaluated: from 2021 to 2023, the median percentage of 'pass' scores per dataset for the minimum standard goes from 78%, to 67%, to 61%, while the median percentage of 'full' scores per dataset for the standard of excellence goes from 29%, 25%, and 13%, respectively. In 2021, the call for papers for the D&B track required the submission of dataset documentation, URL for accessing the dataset, details about data licensing, hosting, and maintenance, and for authors to ensure easy reproducibility [66]. In the following years, additional requirements around datasets being in widely used formats, long-term preservation, inclusion of and access to metadata, and usage of persistent identifiers were added [69, 72]. Despite the increasing stringency of requirements, we could not find any evidence in this sample to suggest that the extent of provided documentation improved over time, pointing to the need for an encompassing structure and framework by which to assess documentation practices and pinpoint areas of improvement. Furthermore, the use of structured documentation also reduced over time, from 80% in 2021 to 50% in 2023.



(a) Minimum standard



(b) Standard of excellence

Figure 3: Temporal distribution across years 2021-2023, **(a)** 'pass scores' for the minimum standard of quality and **(b)** 'full scores' for the standard of excellence across elements. In both cases there is no change across time; see **R6**.

5 Discussion

5.1 How to Improve Data Curation at NeurIPS: Strategies and Resources

Our findings identify areas for which datasets have low or no documentation. To aid dataset creators in strengthening their curation processes, we recommend the adoption of the following methods and resources, particularly for the elements 'requirements', 'ethicality', 'context awareness', 'environmental footprint', 'findability', and 'reusability'.

First, to address **'requirements'** we echo recommendations regarding the creation of **purpose statements** [4]. Stating how dataset creators translated the "real-world" problem they are addressing into a "ML problem" for which the dataset is created [64, 79] promotes transparency. This process consists of numerous decisions, expertise, and assumptions that should be documented in order to understand the context in which the problem situation was framed. In cases of harmful ML models, it has been seen that this translation process can lead to the creation of bad proxy data that unfairly represents the real-world scenario [77]. Furthermore, sharing this process reveals the practicalities of performing data work which is often hidden and considered under-valued [33, 95]. Particularly, it is important for dataset creators to distinguish between the **initial formulation** of the problem **vs.** the **dataset creation scheme** detailing how the dataset development was actually executed. The latter is often documented after the development of the data and can be impaired by the "forgetting practice" of only recording conclusions [64]. As Muller et al. distinctly point out, "Measurement plans tend to record conclusions, not rationales… Other people then work with those conclusions, and have no way to access those unrecorded rationales." [64, p. 9].

We urge dataset creators to explicitly document how the benefits of developing their dataset outweigh the harms of creating it to improve reflection on **'ethicality'**. In other words, the **proportionality principle** must be considered. In ethics, it is understood that actions have positive and negative effects simultaneously. This is called the double effect. "Applications of double effect always presuppose that some kind of proportionality condition has been satisfied. Traditional formulations of the proportionality condition require that the value of promoting the good end outweigh the disvalue of the harmful side effect." [59]. The submission checklist for NeurIPS requires authors to document potential positive and negative societal impacts of their work; we support this and additionally recommend the checklist be amended to encourage comparison and reflection on these in proportion to each other, which very few datasets from our evaluations explicitly do.

'Context awareness' was not well demonstrated in any dataset we evaluated, correlated with the lack of positionality statements in the NeurIPS D&B track as a whole. Past research from the track has pointed to the importance of annotator **positionality**, as researchers' social identity and implicit biases impact data-related choices [4]; this recommendation is also made from a feminist

HCI (human-computer interaction) perspective with the goal to increase **reflexivity** in ML dataset development [91]. See Appendix for examples of positionality and reflection statements of this work.

Data curation in ML encompasses many phases, each responsible for significant emissions due to energy consumption [21, 36, 45, 55, 76, 80]. Several datasets in our sample ranged from over one million instances [37, 41, 89, 100] to several billion instances [22, 28, 32, 83]. Owing to their volume, these datasets are anticipated to have a significant environmental footprint, beyond the impacts associated with model training which have tended to be the focus in ML. There were no quantitative assessments of **'environmental footprint'** in the evaluated datasets. This quantification is crucial for several reasons: it allows for the assessment and comparison of carbon footprint across different projects, facilitating more informed decisions about resource allocation and model design [55, 94]. Understanding these impacts can also drive the development of more energy-efficient algorithms and hardware, contributing to broader efforts to mitigate climate change [40, 45, 96]. Moreover, as public awareness of environmental issues grows, the ML community faces increasing pressure to demonstrate accountability and progress toward sustainability [40, 80, 94]. Transparent reporting of carbon emissions can enhance the credibility of research institutions and companies in the field. By understanding where these emissions originate, researchers and engineers can better target interventions, such as optimizing algorithms for efficiency or sourcing providers with renewable energy for power-intensive tasks, to mitigate the ecological consequences of their work [40, 45, 96]. We provide a list of strategies in the Appendix on how to report environmental footprint for dataset development.

To improve '**findability**', we urge NeurIPS to require datasets to have **metadata** (not just data) assigned a persistent identifier and hosted in a searchable repository (such as Zenodo). While we recognize that having this requirement for all datasets may be infeasible due to volume, sensitivity, and other factors, having this information *for metadata* will help provide information about the dataset that will be available in the long-term, even if the data are gone. Although many datasets are provided on GitHub or platforms hosted by the dataset creators, these URLs suffer from a high likelihood of **link rot** (that the link will no longer be available over time) [42]. Lastly, to improve '**reusability**', we echo the inclusion of identifier information, dataset characteristics, and dataset provenance as outlined in The Data Provenance Initiative [54].

5.2 What Next: A Proposal for Peer-Review

Peer-review processes are constantly progressing with evolving needs, and they require multiple lenses [14, 15]. We propose that our evaluation framework can provide a structure that enhances the submission and peer-review process for the NeurIPS Datasets and Benchmarks track. Dataset creators, by applying the rubric and associated resources, would be more easily able to conduct their dataset development processes with data curation in mind. On the other side, dataset reviewers can use the rubric to evaluate dataset documentation, highlight targeted areas of improvement where data curation standards can be applied, and provide recommendations. The framework speaks directly to the evaluation of documentation required from peer-reviewers in the review form - "Documentation: For datasets, is there sufficient detail on data collection and organization, availability and maintenance, and ethical and responsible use? Note that dataset submissions should include documentation and intended uses; a URL for reviewer access to the dataset; and a hosting, licensing and maintenance plan." [70]. The presented framework allows for the systematic, precise, and encompassing evaluation of these details beyond the prompts present in the current review form, through the criteria presented for the elements in the data pipeline category (i.e., data collection and organization), the data management category based on the FAIR principles [99] (i.e., availability and maintenance), and the ethicality and reflexivity category (i.e., ethical and responsible use). The use of this framework would also enable reviewers to consistently determine whether an additional ethics review is needed. A past consistency experiment for peer-review at NeurIPS in 2014 showed that "if the conference reviewing had been run with a different committee, only half of the papers presented at the conference would have been the same" [20, p. 3]. The 2021 version of the experiment was "...consistent with the 2014 experiment when the conference was an order of magnitude smaller." [9]. Our framework can thus help provide a means of consistency in terms of dataset documentation. We suggest incorporating the framework by introducing a dedicated 'dataset documentation' reviewer role to the NeurIPS D&B track. This can initially be similar to the ethics reviewer, who performs reviews only for those papers that are flagged, but may later evolve to be a part of the core peer-review team. We recommend a dedicated reviewer because of the relatively intense process of evaluating each dataset and the specialized skillset that it will require. Although we found that the time requirement to evaluate each dataset ranged from 35 minutes to 2 hours, the average time in later rounds was limited to 1 hour. Additionally, as with ethics reviews, the results from the dataset documentation review should advise the acceptance of the paper as poor documentation ultimately leads to poor reusability and reproducibility which would defeat the purpose of the D&B track.

5.3 Limitations

We identify five limitations of our work. **1.** The results we showcase are based on the sample set of datasets we evaluated. Although these datasets were randomly chosen and evenly distributed between 2021-2023, there is a chance for the selected datasets to be unrepresentative of the datasets and benchmarks published at the NeurIPS D&B track as a whole. **2.** Our analysis is based on descriptive and interpretative evaluations completed by a mix of reviewers. Although we took careful steps in our iterative process to clarify and normalize standards of interpretation, all results are contingent on the human processes of differing evaluation styles. As with all peer-review, the results are thus dependent on the reviewers. **3.** It is understood that a given reviewer would evaluate each element in the rubric similarly across all datasets. However, we also conducted a disagreement review. This means that a reviewer could change their perspective for a specific element based on the comments of another reviewer, if a disagreement was flagged. This does *not* mean that the reviewer would then update their scores and comments for that element for all *other* datasets they evaluated. Thus some inconsistencies in how the one element is evaluated across all the datasets might be introduced as the cost of improving within-dataset review quality. We believe the impact of this limitation is low, as our results showed minimal disagreement in the final round (see Figure 1a). **4.** Our rubric is designed to enable qualitative and holistic evaluation of each paper on a standardized basis. We believe strongly in the merits of this approach, however it does limit the amount of insight we can have about how properties of the dataset itself influence curation practices. An interesting complementary future approach to study this would be to code characteristics of datasets and see if the trends we identify vary when decomposed by these characteristics, i.e., whether there are specific documentation trends across types of ML datasets or metadata. **5.** Using documentation to understand the curation process is not a substitute for being directly involved in the process or communicating with the dataset creators. As such, it is possible and ultimately a limitation that the reality of data curation is more complex than what is covered in documentation or our framework. In such cases, things can become overly simplified in the documentation and auditing process, e.g., box ticking instead of genuine reflection and evaluation. An ethnographic study of data curation would yield different, additional, insights that we cannot provide. This is currently a limitation and opportunity for future work.

6 Conclusion

By giving datasets and benchmarks a dedicated venue, NeurIPS has sent a clear message that dataset quality is the foundation of continued progress in ML applications and fundamental understanding. There is no better database of knowledge than data curation to aid in this venture. Our evaluation framework adopts concepts from these fields for ML and provides a practical lens on how NeurIPS can spearhead the requirement for rigorous data curation in ML. The enhancements due to the framework are designed to improve the quality, ethicality, and human oversight of new datasets and benchmarks, fostering greater scientific integrity and advancement.

Acknowledgments and Disclosure of Funding

This research was partially supported by NSERC through RGPIN-2016-06640 and by the Canada Foundation for Innovation and the Ontario Research Fund.

References

[1] Stephen Abrams. 2015. A foundational framework for digital curation: The Sept domain model. In *iPRES 2015, The 12th International Conference on Digital Preservation*. http://escholarship.org/uc/item/75v3z67n.pdf

[2] AIES. [n. d.]. Call for Papers AIES 2024. https://www.aies-conference.com/2024/call-for-papers/

[3] Saad Almohaimeed, Saleh Almohaimeed, Ashfaq Ali Shafin, Bogdan Carbunar, and Ladislau Bölöni. 2023. THOS: A Benchmark Dataset for Targeted Hate and Offensive Speech. In *Workshop on Data-centric machine learning*.

[4] Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. 2023. Ethical Considerations for Responsible Data Curation. Advances in Neural Information Processing Systems.

[5] Adrian Arnaiz-Rodriguez and Nuria Oliver. 2024. Towards Algorithmic Fairness by Means of Instance-Level Data Re-Weighting Based on Shapley Values. In *Workshop on Data-centric machine learning*.

[6] Devansh Arpit, Stanisław Jastrzundefinedbski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 233–242.

[7] Julia Barnett. 2023. The Ethical Implications of Generative Audio Models: A Systematic Literature Review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 146–161. https://doi.org/10.1145/3600211.3604686

[8] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041 Place: Cambridge, MA Publisher: MIT Press.

[9] Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021. The NeurIPS 2021 Consistency Experiment. https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment/

[10] Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, and Christoph Becker. 2024. Dataset for "The State of Data Curation at NeurIPS: An Assessment of Dataset Development Practices in the Datasets and Benchmarks Track". https://doi.org/10.5281/zenodo.11398627

[11] Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, and Christoph Becker. 2024. Machine Learning Data Practices through a Data Curation Lens: An Evaluation Framework. In *2024 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3630106.3658955 arXiv:2405.02703 [cs].

[12] Christine L. Borgman. 2015. *Big data, little data, no data: scholarship in the networked world*. The MIT Press, Cambridge, Massachusetts.

[13] Carolyn Ashurst, Markus Anderljung, Carina Prunkl, Jan Leike, Yarin Gal, Toby Shevlane, and Allan Dafoe. 2020. A Guide to Writing the NeurIPS Impact Statement. https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832

[14] NeurIPS Communication Chairs. 2021. NeurIPS 2021: Changes to the Review Process. https://blog.neurips.cc/2021/04/09/neurips-2021-changes-to-the-review-process/

[15] NeurIPS Communication Chairs. 2021. A Retrospective on the NeurIPS 2021 Ethics Review Process. `https://blog.neurips.cc/2021/12/03/a-retrospective-on-the-neurips-2021-ethics-review-process/`

[16] Tiffany C. Chao, Melissa H. Cragin, and Carole L. Palmer. 2014. Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. *Journal of the Association for Information Science and Technology* (June 2014), n/a–n/a. `https://doi.org/10.1002/asi.23184`

[17] Valeriia Cherepanova, Steven Reich, Samuel Dooley, Hossein Souri, John Dickerson, Micah Goldblum, and Tom Goldstein. 2023. A Deep Dive into Dataset Imbalance and Bias in Face Identification. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 229–247. `https://doi.org/10.1145/3600211.3604691`

[18] Kasia S. Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2022. The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence. In *NeurIPS 2020 Workshop on Dataset Curation and Security*. `http://arxiv.org/abs/2201.03954`

[19] Giovanni Colavizza, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf. 2022. Archives and AI: An Overview of Current Debates and Future Perspectives. *Journal on Computing and Cultural Heritage* 15, 1 (Feb. 2022), 1–15. `https://doi.org/10.1145/3479010`

[20] Corinna Cortes and Neil D. Lawrence. 2021. Inconsistency in Conference Peer Review: Revisiting the 2014 NeurIPS Experiment. (2021). `https://doi.org/10.48550/arXiv.2109.09774` arXiv:2109.09774 [cs].

[21] Miyuru Dayarathna, Yonggang Wen, and Rui Fan. 2015. Data center energy consumption modeling: A survey. *IEEE Communications surveys & tutorials* 18, 1 (2015), 732–794. Publisher: IEEE.

[22] Melissa Dell, Jacob Carlson, Tom Bryan, Emily Silcock, Abhishek Arora, Zejiang Shen, Luca D'Amico-Wong, Quan Le, Pablo Querubin, and Leander Heldring. 2023. American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers. Advances in Neural Information Processing Systems.

[23] Junwei Deng and Jiaqi Ma. 2024. Computational Copyright: Towards a Royalty Model for AI Music Generation Platforms. In *Workshop on Data-centric machine learning*.

[24] Advait Deshpande and Helen Sharp. 2022. Responsible AI Systems: Who are the Stakeholders?. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 227–236. `https://doi.org/10.1145/3514094.3534187`

[25] DMLR. [n. d.]. Call for Papers DMLR 2024. `https://dmlr.ai/cfp-icml24/`

[26] Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. 2021. On the Role of Data in PAC-Bayes Bounds. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 130)*, Arindam Banerjee and Kenji Fukumizu (Eds.). PMLR, 604–612. `https://proceedings.mlr.press/v130/karolina-dziugaite21a.html`

[27] Gintare Karolina Dziugaite and Daniel M Roy. 2017. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *Association for Uncertainty in Artificial Intelligence* (2017).

[28] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, and others. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems* 36 (2024).

[29] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92. `https://doi.org/10.1145/3458723`

[30] Rafael Mosquera Gómez, Julian Eusse, Juan Ciro, Daniel Galvez, Ryan Hileman, Kurt Bollacker, and David Kanter. 2023. Speech Wikimedia: A 77 Language Multilingual Speech Dataset. In *Workshop on Data-centric machine learning*.

[31] Kevin A. Hallgren. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology* 8, 1 (2012), 23–34. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/

[32] Eric Hambro, Roberta Raileanu, Danielle Rothermel, Vegard Mella, Tim Rocktäschel, Heinrich Küttler, and Naila Murray. 2022. Dungeons and data: A large-scale nethack dataset. *Advances in Neural Information Processing Systems* 35 (2022), 24864–24878.

[33] Amy K. Heger, Liz B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–29. https://doi.org/10.1145/3555760

[34] Sarah Higgins. 2008. The DCC Curation Lifecycle Model. *International Journal of Digital Curation* 3, 1 (2008), 134–140. https://doi.org/10.2218/ijdc.v3i1.48

[35] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proc. CHI'2019*. ACM, Glasgow Scotland Uk, 1–16. https://doi.org/10.1145/3290605.3300830

[36] Energy Innovation. 2020. How Much Energy Do Data Centers Really Use? https://energyinnovation.org/2020/03/17/how-much-energy-do-data-centers-really-use/

[37] Md Mofijul Islam, Reza Mirzaiee, Alexi Gladstone, Haley Green, and Tariq Iqbal. 2022. CAESAR: An embodied simulator for generating multimodal referring expression datasets. *Advances in Neural Information Processing Systems* 35 (2022), 21001–21015.

[38] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. 2020. Fantastic Generalization Measures and Where to Find Them. *International Conference on Learning Representations (ICLR)* (2020).

[39] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 306–316. https://doi.org/10.1145/3351095.3372829

[40] Lynn H Kaack, Priya L Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick. 2022. Aligning artificial intelligence with climate change mitigation. *Nature Climate Change* 12, 6 (2022), 518–527. Publisher: Nature Publishing Group UK London.

[41] Julia Kaltenborn, Charlotte Lange, Venkatesh Ramesh, Philippe Brouillard, Yaniv Gurwicz, Chandni Nagda, Jakob Runge, Peer Nowack, and David Rolnick. 2023. ClimateSet: A large-scale climate model dataset for machine learning. *Advances in Neural Information Processing Systems* 36 (2023), 21757–21792.

[42] Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLOS ONE* 9, 12 (2014), e115253. https://doi.org/10.1371/journal.pone.0115253

[43] Bernard Koch, Emily Denton, Alex Hanna, and Jacob Gates Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. Advances in Neural Information Processing Systems.

[44] Ravin Kohli, Matthias Feurer, Katharina Eggensperger, Bernd Bischl, and Frank Hutter. 2024. Towards Quantifying the Effect of Datasets for Benchmarking: A Look at Tabular Machine Learning. In *Workshop on Data-centric machine learning*.

[45] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700* (2019).

[46] Bianca Lamm and Janis Keuper. 2024. Retail-786k: a Large-Scale Dataset for Visual Entity Matching. In *Workshop on Data-centric machine learning*. https://doi.org/10.48550/arXiv.2309.17164

[47] Susan Leavy, Eugenia Siapera, and Barry O'Sullivan. 2021. Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race. In *Proc. of 2021 AAAI/ACM Conf. on AI, Ethics, and Society*. ACM, Virtual Event USA, 695–703.

[48] Alycia Lee, Brando Miranda, Sudharsan Sundar, Allison Casasola, and Sanmi Koyeyo. 2024. Beyond Scale: The Diversity Coefficient as a Data Quality Metric for Variability in Natural Language Data. In *Workshop on Data-centric machine learning*.

[49] Christopher A. Lee and Helen Tibbo. 2011. Where's the Archivist in Digital Curation? Exploring the Possibilities through a Matrix of Knowledge and Skills. *Archivaria* 72, 0 (Dec. 2011), 123–168. http://archivaria.ca/index.php/archivaria/article/view/13362

[50] Nianyun Li, Naman Goel, and Elliott Ash. 2022. Data-Centric Factors in Algorithmic Fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Oxford United Kingdom, 396–410. https://doi.org/10.1145/3514094.3534147

[51] Yuanzhi Li and Yingyu Liang. 2018. Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/54fe976ba170c19ebae453679b362263-Paper.pdf

[52] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, L. Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. 2022. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence* 4, 8 (2022), 669–677. https://doi.org/10.1038/s42256-022-00516-1

[53] David Liu, Priyanka Nanayakkara, Sarah Ariyan Sakha, Grace Abuhamad, Su Lin Blodgett, Nicholas Diakopoulos, Jessica R. Hullman, and Tina Eliassi-Rad. 2022. Examining Responsibility and Deliberation in AI Impact Statements and Ethics Reviews. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 424–435. https://doi.org/10.1145/3514094.3534155

[54] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2023. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. http://arxiv.org/abs/2310.16787 arXiv:2310.16787 [cs].

[55] Alexandra Sasha Luccioni and Alex Hernandez-Garcia. 2023. Counting carbon: A survey of factors influencing the emissions of machine learning. *arXiv preprint arXiv:2302.08476* (2023).

[56] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376445

[57] Kenneth O. McGraw and S. P. Wong. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1, 1 (1996), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

[58] Reid McIlroy-Young, Jon Kleinberg, Siddhartha Sen, Solon Barocas, and Ashton Anderson. 2022. Mimetic Models: Ethical Implications of AI that Acts Like You. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 479–490. https://doi.org/10.1145/3514094.3534177

[59] Alison McIntyre. 2023. Doctrine of Double Effect. https://plato.stanford.edu/archives/win2023/entries/double-effect/

[60] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–14. https://doi.org/10.1145/3492853

[61] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 161–172. https://doi.org/10.1145/3442188.3445880

[62] Carlos Mougan, Richard Plant, Clare Teng, Marya Bazzi, Alvaro Cabrejas-Egea, Ryan Sze-Yin Chan, David Salvador Jasin, Martin Stoffel, Kirstie Jane Whitaker, and Jules Manser. 2023. How to Data in Datathons. Advances in Neural Information Processing Systems.

[63] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–15. https://doi.org/10.1145/3290605.3300356

[64] Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3517644

[65] NeurIPS. [n. d.]. NeurIPS Code of Ethics. https://nips.cc/public/EthicsGuidelines

[66] NeurIPS. 2021. Call For Datasets and Benchmarks - NeurIPS. https://web.archive.org/web/20210407213644/https://neurips.cc/Conferences/2021/CallForDatasetsBenchmarks

[67] NeurIPS. 2021. NeurIPS 2021 Fact Sheet. https://neurips.cc/media/Press/NeurIPS_2021-Fact_Sheet.pdf

[68] NeurIPS. 2021. NeurIPS 2021 Paper Checklist Guidelines. https://neurips.cc/public/guides/PaperChecklist

[69] NeurIPS. 2022. Call For Datasets and Benchmarks - NeurIPS. https://web.archive.org/web/20220521130452/https://neurips.cc/Conferences/2022/CallForDatasetsBenchmarks

[70] NeurIPS. 2023. https://neurips.cc/Conferences/2023/DatasetsAndBenchmarks/ReviewGuidelines

[71] NeurIPS. 2023. 2023 Ethics Guidelines for Reviewers. https://nips.cc/Conferences/2023/EthicsGuidelinesForReviewers

[72] NeurIPS. 2023. Call For Datasets and Benchmarks - NeurIPS. https://web.archive.org/web/20230325021706/https://neurips.cc/Conferences/2023/CallForDatasetsBenchmarks

[73] NeurIPS. 2023. NeurIPS 2023 Fact Sheet. https://media.neurips.cc/Conferences/NeurIPS2023/NeurIPS2023-Fact_Sheet.pdf

[74] NeurIPS. 2024. Neural Information Processing Systems Code Of Conduct. https://neurips.cc/public/CodeOfConduct

[75] Daniel Noonan and Tamar Chute. 2014. Data Curation and the University Archives. *The American Archivist* 77, 1 (2014), 201–240. https://doi.org/10.17723/aarc.77.1.m49r46526847g587

[76] Office of Energy Efficiency & Renewable Energy. 2023. Data Centers and Servers: Buildings. https://energy.gov/eere/buildings/data-centers-and-servers

[77] Cathy O'Neil. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

[78] Carole L. Palmer, Nicholas Weber, Trevor Muñoz, and Allen Renear. 2013. Foundations of Data Curation: The Pedagogy and Practice of "Purposeful Work" with Research Data. *Archive Journal* 3 (2013). http://www.archivejournal.net/issue/3/archives-remixed/foundations-of-data-curation-the-pedagogy-and-practice-of-purposeful-work-with-research-da

[79] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3287560.3287567

[80] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350* (2021).

[81] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336. https://doi.org/10.1016/j.patter.2021.100336

[82] Rohith Peddi, Shivvrat Arya, Bhrath Challa, Likhitha Pallapothula, Akshay Vyas, Qifan Zhang, Jikai Wang, Vasundhara Komaragiri, Eric Ragan, Nicholas Ruozzi, Yu Xiang, and Vibhav Gogate. 2023. Put on your detective hat: What's wrong in this video?. In *Workshop on Data-centric machine learning*.

[83] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116* (2023).

[84] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers. Advances in Neural Information Processing Systems.

[85] Aabha Pingle, Aditya Vyawahare, Isha Joshi, Rahul Tangsali, and Raviraj Joshi. 2023. L3Cube-MahaSent-MD: A Multi-domain Marathi Sentiment Analysis Dataset and Transformer Models. In *Workshop on Data-centric machine learning*.

[86] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1776–1826. https://doi.org/10.1145/3531146.3533231

[87] Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. Advances in Neural Information Processing Systems.

[88] Seamus Ross. 2012. Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries. *New Review of Information Networking* 17, 1 (2012), 43–68. https://doi.org/10.1080/13614576.2012.679446

[89] Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. 2020. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. *arXiv preprint arXiv:2008.07146* (2020).

[90] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445518

[91] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–37. https://doi.org/10.1145/3476058

[92] William Seymour, Xiao Zhan, Mark Coté, and Jose Such. 2023. A Systematic Review of Ethical Concerns with Voice Assistants. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 131–145. https://doi.org/10.1145/3600211.3604679

[93] Rajat Shinde, Sujit Roy, Christopher E Phillips, Aman Gupta, Aditi Sheshadri, Manil Maskey, and Rahul Ramachandran. 2024. WINDSET: Weather Insights and Novel Data for Systematic Evaluation and Testing. In *Workshop on Data-centric machine learning*.

[94] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).

[95] Andrea K. Thomer, Dharma Akmon, Jeremy J. York, Allison R. B. Tyler, Faye Polasek, Sara Lafia, Libby Hemphill, and Elizabeth Yakel. 2022. The Craft and Coordination of Data Curation: Complicating Workflow Views of Data Science. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 414:1–414:29. https://doi.org/10.1145/3555139

[96] Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. 2020. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558* (2020).

[97] Nanna Bonde Thylstrup. 2022. The ethics and politics of data sets in the age of machine learning: deleting traces and encountering remains. *Media, Culture & Society* 44, 4 (May 2022), 655–671. https://doi.org/10.1177/01634437211060226

[98] Artem Vysogorets and Julia Kempe. 2024. Towards Robust Data Pruning. In *Workshop on Data-centric machine learning*.

[99] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (2016), 160018. https://doi.org/10.1038/sdata.2016.18

[100] Huang Xuanwen, Yang Yang, Wang Yang, Wang Chunping, Zhang Zhisheng, Xu Jiarong, Chen Lei, and Vazirgiannis Michalis. 2022. DGraph: A Large-Scale Financial Dataset for Graph Anomaly Detection. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.

[101] Rubing Yang, Jialin Mao, and Pratik Chaudhari. 2022. Does the Data Induce Capacity Control in Deep Learning?. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 25166–25197. https://proceedings.mlr.press/v162/yang22k.html

[102] Serena Yeung and Joaquin Vanschoren. 2021. Announcing the NeurIPS 2021 Datasets and Benchmarks Track. https://neuripsconf.medium.com/announcing-the-neurips-2021-datasets-and-benchmarks-track-644e27c1e66c

[103] Dorothy Zhao, Alice Xiang, Jerone T A Andrews, and Orestis Papakyriakopoulos. 2024. Measuring Diversity in Datasets. In *Workshop on Data-centric machine learning*.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim that assessing dataset development from a data curation lens can improve the documentation practices in the NeurIPS Datasets and Benchmarks track. We contribute an evaluation framework to support this claim. The abstract and introduction provide further details on our paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We outline the limitations of our methods and resulting findings in Section 5.3.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We do not perform experiments in the traditional sense, but we provide methods to reproduce the results provided in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper does not include experiments. However the rubric evaluations and analyses are hosted on Zenodo.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our work does not present a dataset or benchmark. Our work provides a framework to increase accountability, transparency, reusability, and reproducibility of ML datasets.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper does not pose any such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce new assets in the form of evaluations of datasets using our framework. The process of performing these evaluations is discussed in Section 3.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.